Apr 13, 2025

Small Language Models Are the New Rage, Researchers Say

Larger models can pull off a wider variety of feats, but the reduced footprint of smaller models makes them attractive tools.

Large language models work well because they're so large. The latest models from OpenAI, Meta, and DeepSeek use hundreds of billions of "parameters"—the adjustable knobs that determine connections among data and get tweaked during the training process. With more parameters, the models are better able to identify patterns and connections, which in turn makes them more powerful and accurate.

But this power comes at a cost. Training a model with hundreds of billions of parameters takes huge computational resources. To train its Gemini 1.0 Ultra model, for example, Google reportedly spent $191 million. Large language models (LLMs) also require considerable computational power each time they answer a request, which makes them notorious energy hogs. A single query to ChatGPT consumes about 10 times as much energy as a single Google search, according to the Electric Power Research Institute.

In response, some researchers are now thinking small. IBM, Google, Microsoft, and OpenAI have all recently released small language models (SLMs) that use a few billion parameters—a fraction of their LLM counterparts.

Small models are not used as general-purpose tools like their larger cousins. But they can excel on specific, more narrowly defined tasks, such as summarizing conversations, answering patient questions as a health care chatbot, and gathering data in smart devices. "For a lot of tasks, an 8 billion–parameter model is actually pretty good," said Zico Kolter, a computer scientist at Carnegie Mellon University. They can also run on a laptop or cell phone, instead of a huge data center. (There's no consensus on the exact definition of "small," but the new models all max out around 10 billion parameters.)

To optimize the training process for these small models, researchers use a few tricks. Large models often scrape raw training data from the internet, and this data can be disorganized, messy, and hard to process. But these large models can then generate a high-quality data set that can be used to train a small model. The approach, called knowledge distillation, gets the larger model to effectively pass on its training, like a teacher giving lessons to a student. "The reason [SLMs] get so good with such small models and such little data is that they use high-quality data instead of the messy stuff," Kolter said.

Researchers have also explored ways to create small models by starting with large ones and trimming them down. One method, known as pruning, entails removing unnecessary or inefficient parts of a neural network—the sprawling web of connected data points that underlies a large model.

Pruning was inspired by a real-life neural network, the human brain, which gains efficiency by snipping connections between synapses as a person ages. Today's pruning approaches trace back to a 1989 paper in which the computer scientist Yann LeCun, now at Meta, argued that up to 90 percent of the parameters in a trained neural network could be removed without sacrificing efficiency. He called the method "optimal brain damage." Pruning can help researchers fine-tune a small language model for a particular task or environment.

For researchers interested in how language models do the things they do, smaller models offer an inexpensive way to test novel ideas. And because they have fewer parameters than large models, their reasoning might be more transparent. "If you want to make a new model, you need to try things," said Leshem Choshen, a research scientist at the MIT-IBM Watson AI Lab. "Small models allow researchers to experiment with lower stakes."

The big, expensive models, with their ever-increasing parameters, will remain useful for applications like generalized chatbots, image generators, and drug discovery. But for many users, a small, targeted model will work just as well, while being easier for researchers to train and build. "These efficient models can save money, time, and compute," Choshen said.

(Original story reprinted with permission from Quanta Magazine, an editorially independent publication of the Simons Foundation whose mission is to enhance public understanding of science by covering research developments and trends in mathematics and the physical and life sciences.)

https://www.wired.com/story/why-researchers-are-turning-to-small-language-models/